# **Title**

"Predictive Data Analytics for Banking Marketing Success using Term Deposit Subscription Classification"

# **Topic Area**

Predictive Data Analysis, Banking, Telemarketing, Finance

# **Research Objectives**

* To develop and compare predictive models (Random Forest, XGBoost, CatBoost, CNN, and TabNet) for classifying term deposit subscriptions using direct marketing campaign data.
* To evaluate the performance of these models based on accuracy, precision, recall, F1-score, and computational efficiency.
* To analyse the most important features on model prediction.
* To determine which model offers the best balance between interpretability and predictive power.

# **Statement of Hypotheses**

* **H1**: Deep learning-based models (CNN, TabNet) will outperform traditional ensemble methods (Random Forest, XGBoost, CatBoost) in predicting term deposit subscriptions, particularly in handling complex feature interactions.
* **H2**: Ensemble models (XGBoost, CatBoost) will provide higher interpretability and generalisability compared to deep learning models, making them more suitable for business use cases where transparency is important.

# **Literature Review**

## **Predictive Analytics in Banking**

Broby (2022) takes a closer look at the use of these techniques in the financial industry where classification, regression, clustering and time series are discussed. The study incorporates these techniques into Decision Support System (DSS), which assists the financial managers in their decisions making. More than 187 papers are inspected according to the proposed SPAR-4-SLR protocol; and despite the fact that the range includes many algorithms, it is dominated by Random Forests, Support Vector Machines (SVM), and Neural Networks. For instance, Huang et al. (2005) employed SVM to anticipate the overall direction of stock at 75%accuracy; Broby (2022). These methods are used in the determination of the financial future position and in the framework of risk management plans.

In the present study, Moinuddin et al. (2024) analyse the effect of the marketing analytics on the consumers and the campaign effectiveness using both discussed interviews and surveys. This research also finds out that checking of analytics tools is done weekly by 75% of the respondents, the mean score the respondents gave to the increase in the success of their campaigns as a result of this activity was 8.5 out of 10. Correlation analysis gave a value of 0.75 on the Pearson coefficient and this suggested that analytics usage had directly dependent on the campaign evaluation (Moinuddin et al., 2024/). This is particularly important in current society, yet to show why there is a need to shift towards data driven marketing.

Zaki et al, (2024) analyse predictive analytics and machine learning in term deposit subscriptions of the bank using models like SGD, KNN, and Random Forest. Feature engineering, cross tabulation, and heatmaps were used during data exploration conducted in the study. Random Forest model performed better than other models where accuracy was found to be 87 %, PPV = 87.83% and NPV = 92.99%. These findings offer relevant tips that can be used to improve advertising tactics in banking (Zaki et al., 2024).

Amajuoyi et al. (2024) study the impact of analytics models for customer retention and business growth. Based on the obtained data of the contacts’ database and machine learning algorithms, they define high and low potential clients and churn threats. According to the study it indicates that the businesses that adopt the actions of predictive analytics experienced a customer retention level of 25% and the efficiency of an expanding market of 30%. The study maps to how machine learning advances targeted customer marketing and interaction (Amajuoyi et al., 2024).

In finance, Olaniyi et al. (2023) consider how predictive analytics shifts data into a form that gives insight. The authors discuss and evaluate various decision approaches including decision tree, random forest, and neural network to show that the predictive models yield the result of 40 percent increase in resource efficiency for financial institutions. The research also focuses on the PWE’s key significance for risk management and fraud prevention; effectiveness enhancement up to 30%, in the essential financial risk situations (Olaniyi et al., 2023).

Most recent work on decision trees for predictive analytics of business decisions include Lee et al. 2022. The study used secondary research where more than twenty four papers have been reviewed from fields as; healthcare, finance and customer relationship. This was especially interesting because the decision tree showed marked improvement of 30% in customer behaviour predictability in relation to their customer retention. Concisely, the authors shall find that although relatively basic and very easy to comprehend, the decision trees are still the valuable instruments for the predictive business analytics (Lee et al., 2022).

Authors Wassouf et al. (2020) implement a big data framework for predicting customer loyalty for the Syriatel Telecom. They used two main components of the TFM model in combination with usage and demographic data of customers to categorise customer loyalty. Among the four machine learning models tested: Random Forest; Decision Tree; Gradient-Boosted Trees; and Multilayer Perceptron. The highest accuracy achieved in binary classification based on the results of the Gradient-Boosted Trees is 87%. This paper also reveals how big data analytics can promote customer loyalty more than it has already been estimated (Wassouf et al., 2020).

Zulaikha et al. (2020) look into how the integration of artificial intelligence in marketing can shape the way customer predictive analytics. By applying various algorithms, the study’s customer data on behaviour and demographic characteristics are segmented and targeted. The authors also note that companies adapting AI within the scheme for customer segmentation saved 25% of the targeting effect and reduced their marketing expenses by 30%. This paper therefore emphasises on the feasibility of the AI based segmentation in improving marketing effectiveness and consumer satisfaction (Zulaikha et al., 2020).

## **Ensemble Models**

Using phone contacts including 45,211 records, Patwary et al. (2021) explore prediction of bank deposit through ensemble learning. Naive Bayes, Support Vector Machine (SVM) and Neural Networks (NN) methods are used in the study, with Bagging, Boosting, and Stacking ensembles. It denotes that the chosen model of Neural Networks (Bagging) displayed higher accuracy of about 96.62% with sensitivity of 97.14% and high specificity of 99.08%. This study points to the effectiveness of ensemble models over the traditional classifiers in predicting the amount which customers would be willing to deposit in banks (Patwary et al., 2021).

Wardani et al. (2023) assess the performance of bias reduction of bank customer data using Machine Learning techniques; XGBoost, LightGBM and Random Forest. This study attempts to use AI Fairness 360 to detect and correct bias using disparate impact (DI) and statistical parity difference (SPD). The experiment follows a mitigation approach and saw a great reduction of bias through adversarial debiasing with a mean DI of 0.943 and SPD of -0.004. The study shows how adversarial debiasing is better than other approaches in arriving at a fairer prediction (Wardani et al., 2023).

Shin et al. (2021) have exhibited and compared Random Forest, GBM, XGBoost, LightGBM decision tree-based ensemble learning technique for predicting term deposit subscriptions. Therefore, in this Kaggle bank marketing dataset study, using 10-fold cross validation, the proposed LightGBM has a higher accurate rate 86.4%, the Kappa coefficient 0.727, F1-score 0.861. In this study, SHAP was used to add interpretability and showed that model decisions were most affected by customer contact duration (Shin et al., 2021).

In the context of a digital marketing dataset, Abd et al. (2024) use Random Forest, Decision Tree, KNN and CatBoost to categorise customer behaviour. CatBoost provided the highest accuracy of %98 and F-measure score %0.91 compared to the Random Forests %97, Decision Trees %95 and KNN %91. While CatBoost took 15.04 seconds to execute, it had an optimal generative predictive performance, recommending CatBoost for use when performing marketing prediction analysis (Abd et al., 2024).

Kinnander (2020) compares the results of the XGBoost, LightGBM, and CatBoost models of Gradient Boosting Tree capabilities in terms of the accuracy of customer profitability prediction in online retail. This work then tests and compares these models based on recall and precision rates, employing the first purchase data of new customers of Klarna. XGBoost and LightGBM were almost equal, with precision of up to 80%, while CatBoost was slightly better in overall performance in classifying profitable customers with a recall of 0.85 (Kinnander, 2020).

In their study, Wankhede et al. (2019) work on enhancing the accuracy of identifying potential clients on the Bank Term Deposits using an explore Machine Learning algorithms that include, but not limited to Logistic Regression, Support Vector Machine (SVM), Random Forest, Neural Networks, and eXtreme Gradient Boosting (XGBoost). Of these Random Forest is the most accurate model with accuracy of 87% after addressing the machine learning classification problem of class imbalance using ‘Spread-subsample’ technique. In general, the study stresses the necessity to derive features for enhancing the predictive success rate of telemarketing (Wankhede et al., 2019).

Nupponen (2024) explores the accuracy of the machine learning models for sales performance of manufacturing service companies. The models used in this study include Logistic Regression, Decision Tree, XGBoost with hyperparameter tuning of XGBoost model. The outcomes show that XGBoost has a better fit than other models; its test classification accuracy reached 84.2% and the AUC was equal to 0.88. Extending the SHAP analysis to hierarchical levels, direct sales channels and service agreements appeared to be the driving force behind successful sales (Nupponen, 2024).

Meshref (2020) uses the Kaggle Bank Marketing dataset to predict loan approvals using ensemble machine learning algorithms including; Bagging, Boosting, LogitBoost, and Random Forests. The models yielded an accuracy of 83.98% with LogitBoost in recognition of the best performance in the models. The study also focused on model interpretability since feature selection and interpreting decision trees were employed; this information was important for banks’ decision making. From this research, Meshref highlighted how increasing the algorithms’ efficiency improves predictive accuracy of the bank marketing dataset (2020).

Chlebus and Osika (2020) analyse five tree-based algorithms, including Random Forest, AdaBoost, GBM, XGBoost, and CatBoost, on the Portuguese bank telemarketing dataset. This research measures model stability and employs XAI tools for a qualitative analysis of the interpretable models. Title: The Performance of XGBoost, CatBoost and LightGBM on the GBDx Breast Cancer Data Set Using Feature Importance With regard to AUC, XGBoost had the best value with a score of 0: 8012, CatBoost had 0: 8003. But when it came to XAI analysis of both algorithms, they noted that while XGBoost was endowed with signs of overfitting, it was wiser to accomplish telemarketing success with CatBoost.

## **Deep Learning**

Dutta et al. (2021) suggest an integration of CNN and GRU for term deposit subscriptions forecast. This study uses telemarketing data from a bank, and measures the performance of the proposed model against benchmark experimental models, such as k-Nearest Neighbour (k-NN), Decision Tree (DT), and Multi-layer Perceptron (MLP). The Convolutional-GRU model has provided an accuracy of 89.59 % and MSE of 0.1041 % higher than other classifiers (Dutta et al., 2021).

In Guha et al. (2024), the authors use machine learning to determine potential customers for term deposits. Based on a dataset of clients of a Portuguese bank, the authors conduct the comparison of Submitted classification algorithms such as SVM, Random Forest, Decision tree and k-NN. Out of these Classifiers, Random Forest reached the highest accuracy of 87%. Making the right customer targeting in banking marketing campaigns can be more efficient using machine learning analysis according to the study of Guha et al., (2024).

Dutta and Bandyopadhyay 2021 developed a hybrid recommender system to predict the probability of subscribers to the term deposit using the)i bank marketing dataset with the help of artificial neural network. The cross validation method used on strata of data gave an overall accuracy of 88.32% an MSE of 0.1168 in the model. When it comes to the models for predicting the probability of the bank term deposit, Dutta and Bandyopadhyay (2021) showed that the neural network model outperformed the ten baseline classifiers, including k-Nearest Neighbours (k-NN), Decision Tree (DT), Multi-layer Perceptron (MLP).

The present study by Clements et al. (2020) introduces Credit Risk Tracking through Deep Learning TCN and LSTM models. The study adopted a sample of 15 million credit card transactions with consideration of early warning on credit risk. TCN model recorded better accuracy of 92.33% in providing the results for outcomes linked with high Recall value of the Gini coefficient of 64.13%. The study acknowledged the fact that deep learning models performed better than GBDT in the financial risk prediction (Clements et al., 2020)

Few feature selection methods chosen and used by Teo and Lee (2019) include k- Best, k – Nearest Neighbour and Closest Centroid for telemarketing success in Portugal for a particular bank after which Decision Tree Model, MLP and Data Randomisation were used. Out of all, Logistic Regression had the highest accuracy of 91.48% while the second accurate model was the Decision Tree with accuracy of 89.91% and MLP having third best accuracy of 90.10%. The study used the CRISP-DM model and showed that its best model of prediction was the Logistic Regression for the level of customer agreement for long-term deposit offers, in line with the findings of Alsolami et al., (2020).

More specifically, Zeinulla et al. (2022) compare classification models for estimating the bank telemarketing success of the Portuguese bank telemarketing dataset. The performance analysis of Random Forest, Deep Artificial Neural Networks (DANN), Logistic Regression, Naive Bayes, k-Nearest Neighbours and Support Vector Machine (SVM) models is done in this study. From the results obtained it was identified that the model with the highest accuracy was Random Forest at 90.88% while DANN had an accuracy of 90.29%. Logistic Regression and Naive Bayes algorithms produced low accuracy as both the results got overfitted from the stage two (Zeinulla et al., 2022).

Dai (2021) also enhances the current Convolutional Neural Network (CNN) in financial forecasting by incorporating data discretisation and Principal Component Analysis (PCA). Through denoise of the dataset and convolutional neural network followed by principal component analysis, the study has obtained rather robust prediction accuracy in the context of financial credit. Analysing empirical Chinese stock market data, it is shown that the updated version of the described CNN model delivered the overall prediction accuracy of 93.2%, which is 12% higher than traditional financial forecasting approaches (Dai, 2021).

Ghatasheh et al. (2020) develop a cost-sensitive Artificial Neural Network (ANN) model for term deposit subscription considering a highly imbalanced exemplary dataset of a Portuguese bank. In order to address the problem of class imbalance the work uses cost-sensitive learning techniques without making any modification to the data. Overall accuracy achieved by using the model was computed on a geometric mean which stood at 79% with Type I and Type II misclassification errors set at 0.192 and 0.229 respectively. This shows that using cost-sensitive analysis can enhance the success rate predictions in telemarketing (Ghatasheh et al., 2020).

For the success prediction of outbound telemarketing for insurance policy loans, Gu et al., (2021 propose an XmCNN. In other words, the model directly employs over 200 elements of the input data as the input indicators with no feature selection, and selects only the feature selection model in the process of controlling overfitting. In the telemarketing success predictors, the proliferation of the proposed XmCNN model outperforms the basic deep learning model in F1-score of 87.47% and FPR of 4.92%. Therefore, the application of the mentioned approach uncovered some of the essential factors that influence or have an impact on telemarketing (Gu et al., 2021).

For estimating the future propensity of deposit customers in the banking industry, Keji, et al. (2024) propose a deep learning model known as Residual Network (ResNet) using Transfer Learning. Pursuing the issue of class imbalance the authors used SMOTE technique and ventured into converting the data set into images for training. The results of the proposed model are compared with the traditional machine learning model in which higher accuracy 93.00%, precision 97.00%, recall 90.00%, and F1 score 93.00% have been achieved. This method provides a very robust method of generating improved forecast solutions for long-term deposit subscriptions (Keji et al., 2024).

Krishna and Reddy (2019) use Deep Neural Networks (DNN) to predict bank marketing data Classification using the following reduction technique: Attribute Subset Selection and Principal Component Analysis. Two datasets of 45,211 and 41,188 instances were used in the study with accuracies of 91.15% and 89.24% gained respectively. The authors also compared the efficiency of the DNN model to general classifiers including Decision Trees, Naïve Bayes, Support Vector Machines, and k-Nearest Neighbours; the results showed the competencies of deep learning in banking marketing predictions were high with greater accuracy (Krishna & Reddy, 2019).

Kim et al., (2015) on the other hand, employ a deep CNN for a telemarketing based on the bank’s success of a sample of the data from Portugal based banking organisations. Thus, by applying deep learning it is possible to obtain 84.9% of accuracy, which is higher than classical methods of machine learning, including decision trees and support vector machines. The same applied to the CNN model, which also demonstrated high effectiveness in the analysis of difficult multi-parametric data obtained from telemarketing campaigns and will be a solid basis for making further marketing predictions in the banking industry in the future (Kim et al., 2015).

In the study of Yüksel (2023), TabNet, a new deep learning architecture, has been evaluated when used for credit risk scoring in comparison to XGBoost and ANN. The study used two datasets for credit risk, which are open source and Table 2 shows that compared to baseline models, TabNet had higher recall of 99.10% while XGBoost had higher accuracy and AUC score. The study also shows that there can be few false negatives with TabNet, which makes the method useful for banks striving for minimising default amount (Yüksel, 2023).

Following machine learning techniques, Rony et al. (2024) analyse the customer behaviour to take long-term bank deposits. In order to fulfil the targets of the study, Exploratory Data Analysis (EDA), Principal Component Analysis (PCA) and five different Machine Learning algorithms namely logistic regression, random forest, support vector machine (SVM), K nearest neighbours and multilayer perceptron (MLP) are used. Based on this criterion, it was found that Logistic Regression belongs to the best of all the algorithms in terms of accuracy 90.64 % and sensitivity 99.05 %. By evaluating the literature, authors attest to the effectiveness of the proposed logistic regression technique for the classification of customers in the banking domain (Rony et al. 2024).

Guo et al. (2023) introduce a novel ensemble approach for predicting the success of bank telemarketing using four artificial neural networks (ANNs) optimised by metaheuristic algorithms: It develops four algorithms – Electromagnetic Field Optimization (EFO), Future Search Algorithm (FSA), Harmony Search Algorithm (HSA) and Social Ski-Driver (SSD). The highest classification rates were obtained from EFO-ANN classification set by AUC 0.8030 and FSA-ANN with AUC of 0.7809 while HSA-ANN obtained a classification rate of AUC 0.7356 only out of 30,488 records. Consequently, the adopted research confirms that ignition of the hybrid metaheuristic models enhances telemarketing performance prediction (Guo et al. 2023).

Kim et al. (2015) consider the success of the bank telemarketing and, to this end, propose the use of a Deep Convolutional Neural Network (DCNN) based on the 45,211 phone calls dataset. This work looks at different hyperparameters in the context of this model such as the number of layers and learning rate for model training. By using the DCNN, the compilation tested the accuracy of 76.70% which is greatly superior to classifiers like logistic regression and support vector machines. Kim et al., (2015) identifies that the research done in this area points to the ability of DCNNs to capture various complexities inherent in financial data for improvement of the predictive qualities.